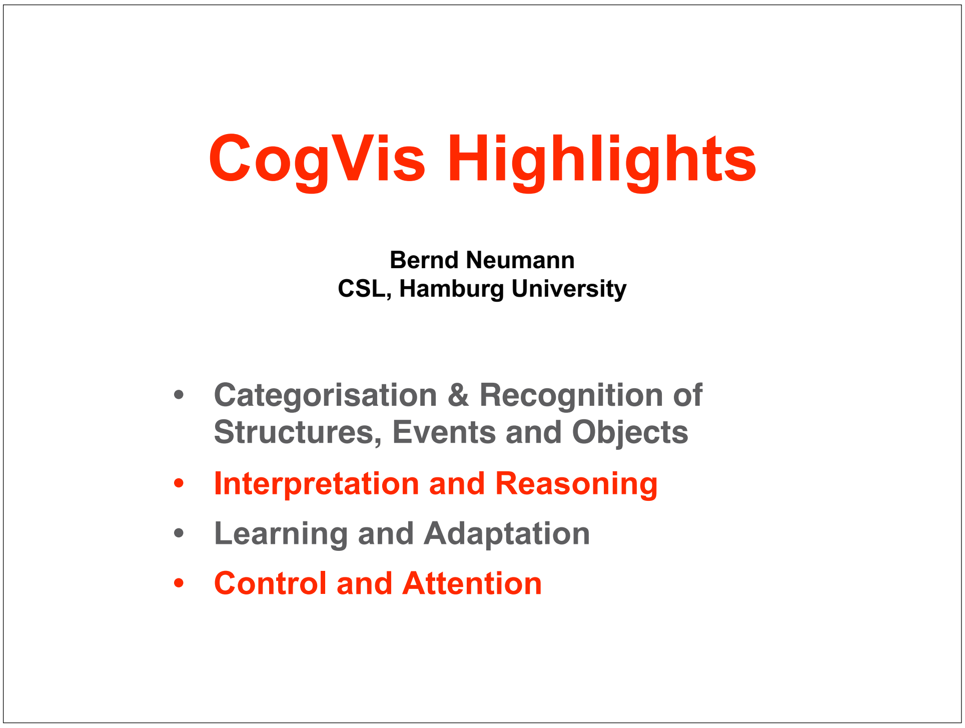


CogVis Highlights

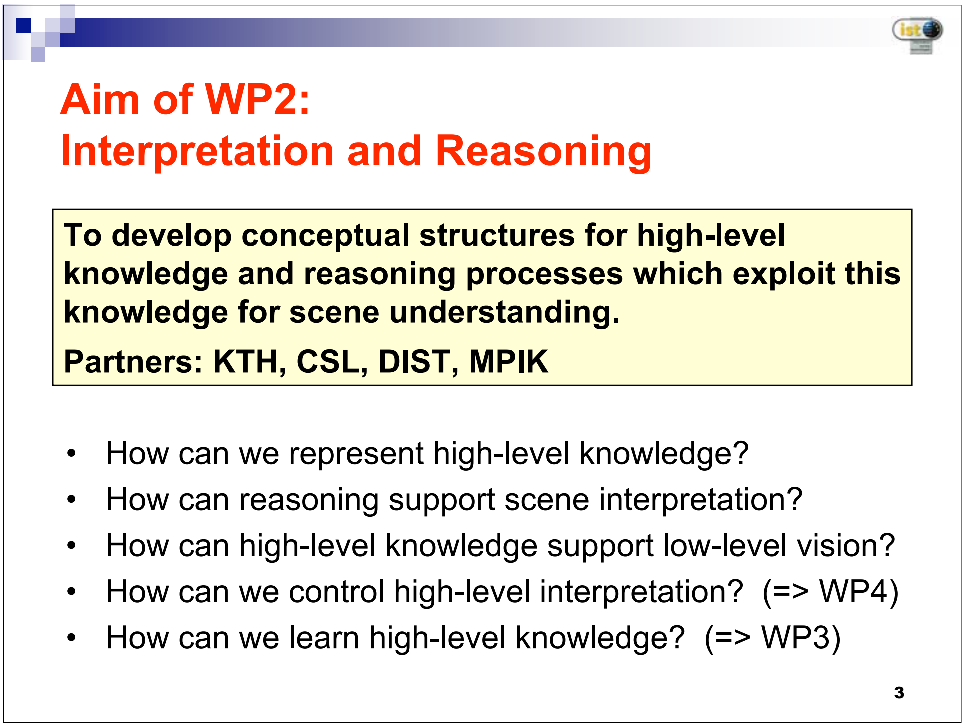
Bernd Neumann
CSL, Hamburg University

- Categorisation & Recognition of Structures, Events and Objects
- Interpretation and Reasoning
- Learning and Adaptation
- Control and Attention

Participants

KTH Stockholm	Henrik I. Christensen,
	Jan-Olof Eklundh
DIST Uni Genua	Julio Sandini
ETH Zürich	Berndt Schiele
LEEDS Uni Leeds	David Hoggs, Tony Cohn
MPIK Tübingen	Heinrich Bülthoff
CSL Hamburg	Bernd Neumann
UOL Ljubljana	Ales Leonardis

2000 - 2004, 56 Person Years, 3.97 Mio Euro

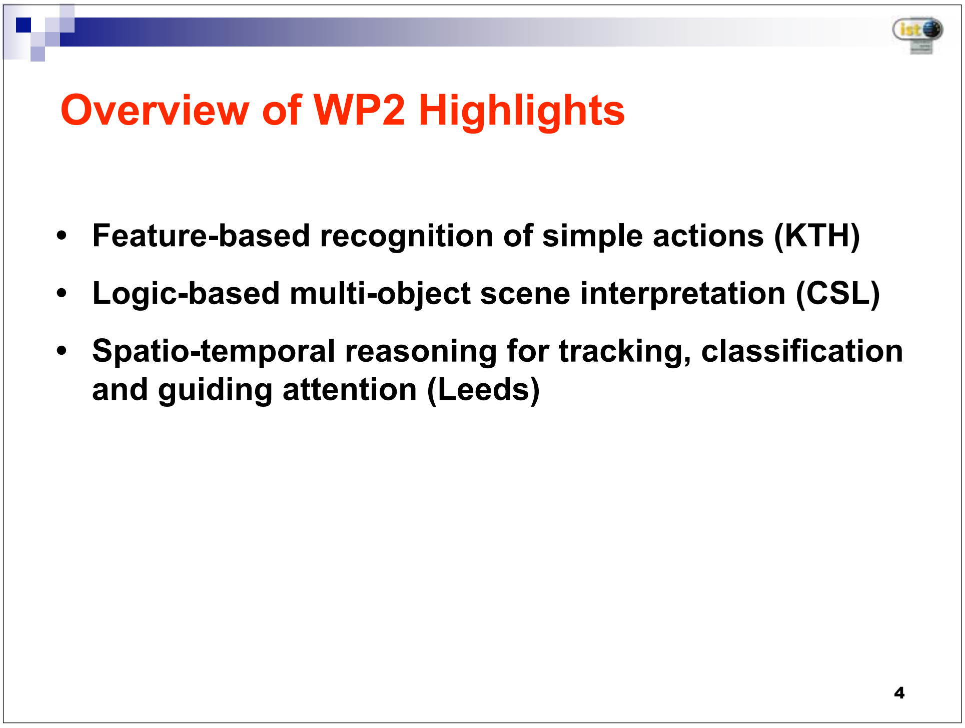


Aim of WP2: Interpretation and Reasoning

To develop conceptual structures for high-level knowledge and reasoning processes which exploit this knowledge for scene understanding.

Partners: KTH, CSL, DIST, MPIK

- How can we represent high-level knowledge?
- How can reasoning support scene interpretation?
- How can high-level knowledge support low-level vision?
- How can we control high-level interpretation? (=> WP4)
- How can we learn high-level knowledge? (=> WP3)



Overview of WP2 Highlights

- **Feature-based recognition of simple actions (KTH)**
- **Logic-based multi-object scene interpretation (CSL)**
- **Spatio-temporal reasoning for tracking, classification and guiding attention (Leeds)**

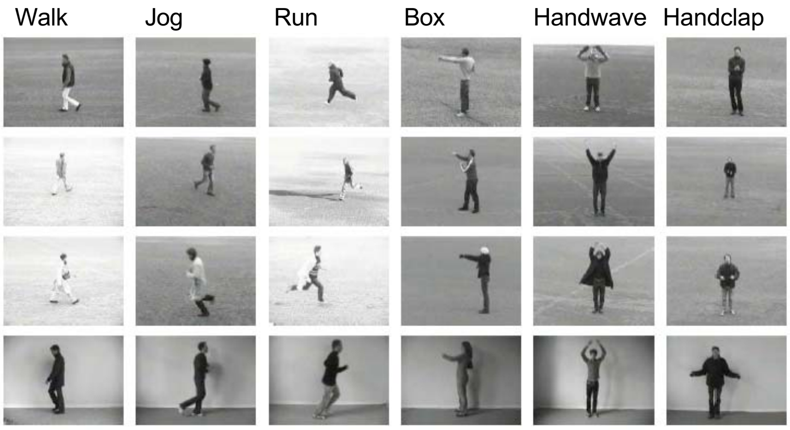
Feature-based Action Recognition

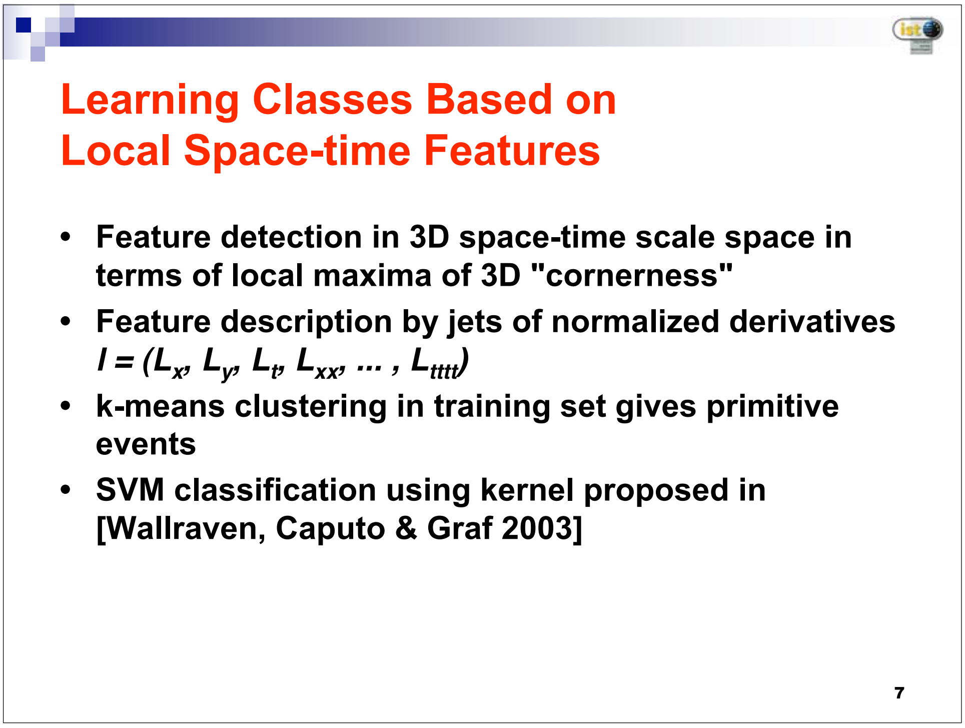
(Schüldt, Caputo & Laptev, KTH)

- **Represent actions by local space-time features** [Laptev & Lindeberg, ICCV03]
- **Learn action classes using SVMs with local kernels** [Wallraven et al., ICCV03]

Action Database

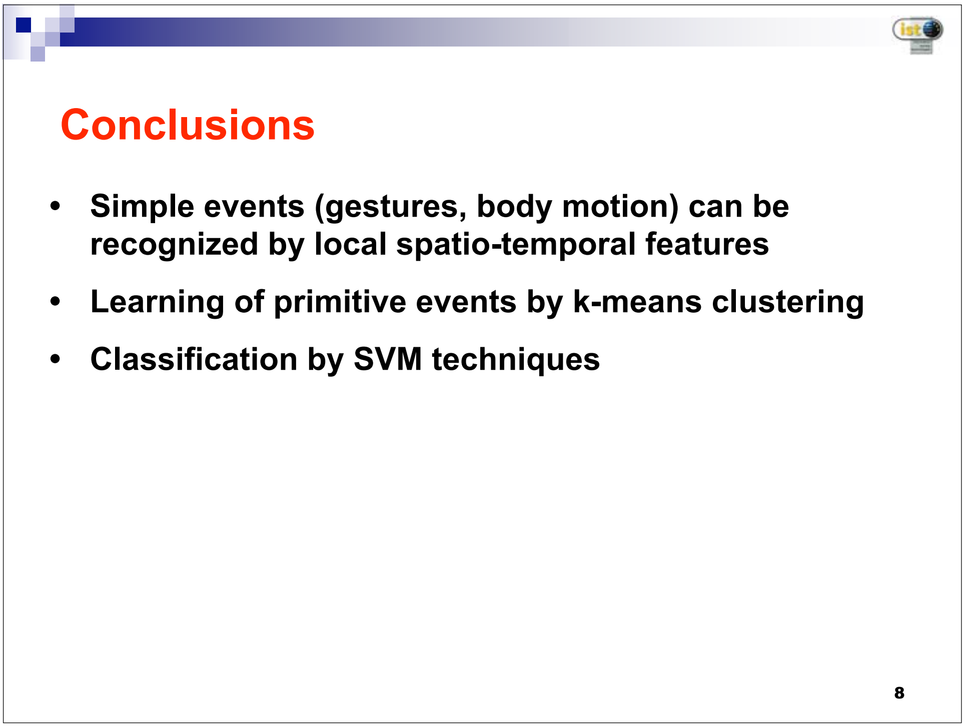
6 actions, 25 subjects, 4 settings, 2400 sequences





Learning Classes Based on Local Space-time Features

- Feature detection in 3D space-time scale space in terms of local maxima of 3D "cornerness"
- Feature description by jets of normalized derivatives $I = (L_x, L_y, L_t, L_{xx}, \ldots, L_{tttt})$
- k-means clustering in training set gives primitive events
- SVM classification using kernel proposed in [Wallraven, Caputo & Graf 2003]



Conclusions

- Simple events (gestures, body motion) can be recognized by local spatio-temporal features
- Learning of primitive events by k-means clustering
- Classification by SVM techniques

Multi-object Scene Interpretation

Differences to object recognition and categorization:

- "Objects" correspond to conceptual units beyond physical objects and at various levels of granularity => compositional hierarchy
- "Objects" may have complex spatio-temporal structure
- Scene understanding may involve common-sense knowledge => taxonomical hierarchy
- Interpretations are typically based on context and partial evidence

Aggregate Structure in a Frame-based Language

```
name:           place-cover
parents:        :is-a agent-activity
parts:          pc-tt :is-a table-top
                pc-tp1 :is-a transport with (tp-obj :is-a plate)
                pc-tp2:is-a transport with (tp-obj :is-a saucer)
                pc-tp3 :is-a transport with (tp-obj :is-a cup)
                pc-cv :is-a cover
time marks:     pc-tb, pc-te :is-a timepoint
constraints:    pc-tp1.tp-ob = pc-cv.cv-pl
                pc-tp2.tp-ob = pc-cv.cv-sc
                pc-tp3.tp-ob = pc-cv.cv-cp
                                ...
                pc-tp3.tp-te ≥ pc-tp2.tp-te
                pc-tb ≤ pc-tp3.tb
                pc-te ≥ pc-cv.cv-tb
```

What is a Scene Interpretation Logically?

Scene interpretation = "partial model"

Model = mapping of formulae into a domain such that all formulae are true

[Reiter & Mackworth 87, Matsuyama & Hwang 90, Schröder 99, Neumann & Möller 04]

Given that low-level vision maps real-world phenomena into correct high-level primitives, then <u>any constructed description</u> which is consistent with evidence, context and generic knowledge is a logically valid scene interpretation.

=> Logics provide a loose frame for scene interpretations

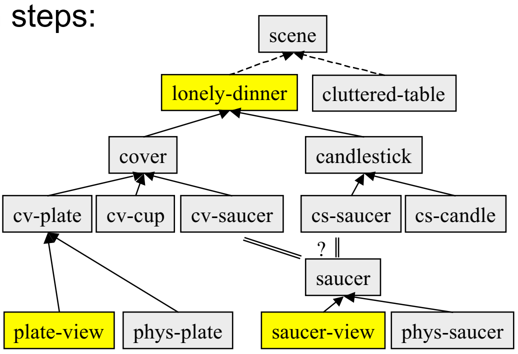
Interpretation as a Stepwise Process

Constructing a description based on compositional and taxonomical concept hierarchies

Four kinds of interpretation steps:

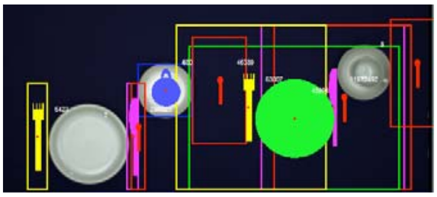
- aggregate instantiation
- instance specialisation
- instance expansion
- instance merging

=> mixed bottom-up and top-down strategies are possible



Experimental Results

natural views = evidence
coloured shapes = hypotheses
boxes = expected locations



Snapshot illustrates intermediate state of interpretation after 89 interpretation steps:

- **hypotheses based on partial evidence**
- **predictions about future actions and locations**
- **high-level disambiguation of low-level classification**
- **influence of context**

Conclusions

- **High-level scene interpretations are modelled by taxonomical and compositional aggregate hierarchies**
- **Description Logics provides an adequate but loose framework for**

  - consistent scene interpretations

  - possible interpretation steps
- **Probabilistic models of aggregates provide**

  - fragments for Bayes Net scene descriptions

  - inference services for preferred interpretations

Enhanced Tracking and Recognition by Enforcing Spatio-Temporal Consistency

(Bennett et al., LEEDS)

Key ideas:

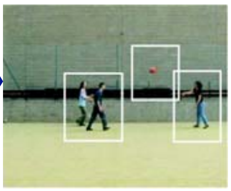
- **Generate models that satisfy spatio-temporal continuity constraints**
- **Select model that is best supported by statistical classifier**
- **Improve labelling accuracy by adding logical reasoning to low level tracking**
- **Increase granularity to enhance reliability**

Processing Steps

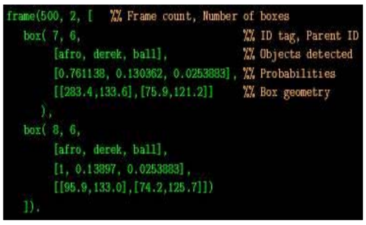
Objects are tracked, boxed and labeled.

Initial low-level processing

Tracker



Classifier



```
frame(500, 2, [   %% Frame count, Number of boxes
  box( 7, 6,                                  %% ID tag, Parent ID
      [afro, derek, ball],                    %% Objects detected
      [0.761138, 0.130362, 0.0253883],        %% Probabilities
      [[283.4,133.6],[75.9,121.2]]            %% Box geometry
    ),
  box( 8, 6,
      [afro, derek, ball],
      [1, 0.13897, 0.0253883],
      [[95.9,133.0],[74.2,125.7]])
  ]).
```

Input to Prolog reasoner

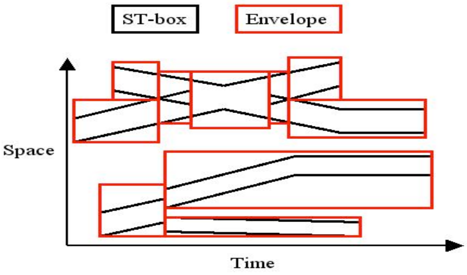
The Spatio-Temporal Consistency Reasoning Algorithm (1)

Envelopes of Constant Occupancy are created.

Overlapping boxes are merged into clusters.

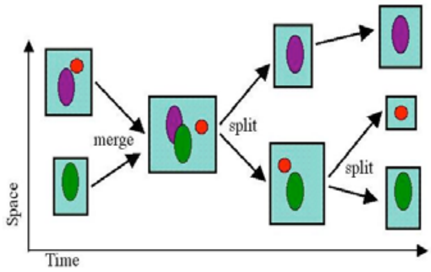
Temporally continuous clusters are aggregated into ***envelopes***, which (to very high probability) have constant occupancy throughout their lives.

Continuity is described by the graph of the parent/child relation between envelopes.



The Spatio-Temporal Consistency Reasoning Algorithm (2)

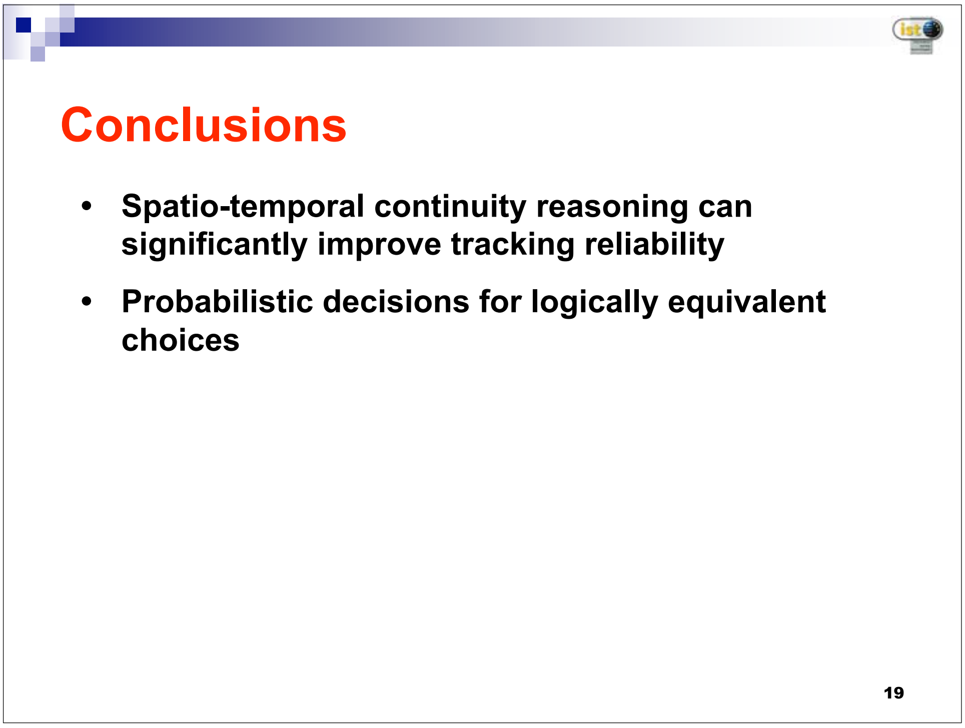
The reasoner progressively generates all spatio-temporally consistent models.



Votes for models are computed based on the statistical classifier output and used to prune the search and find the best supported model.

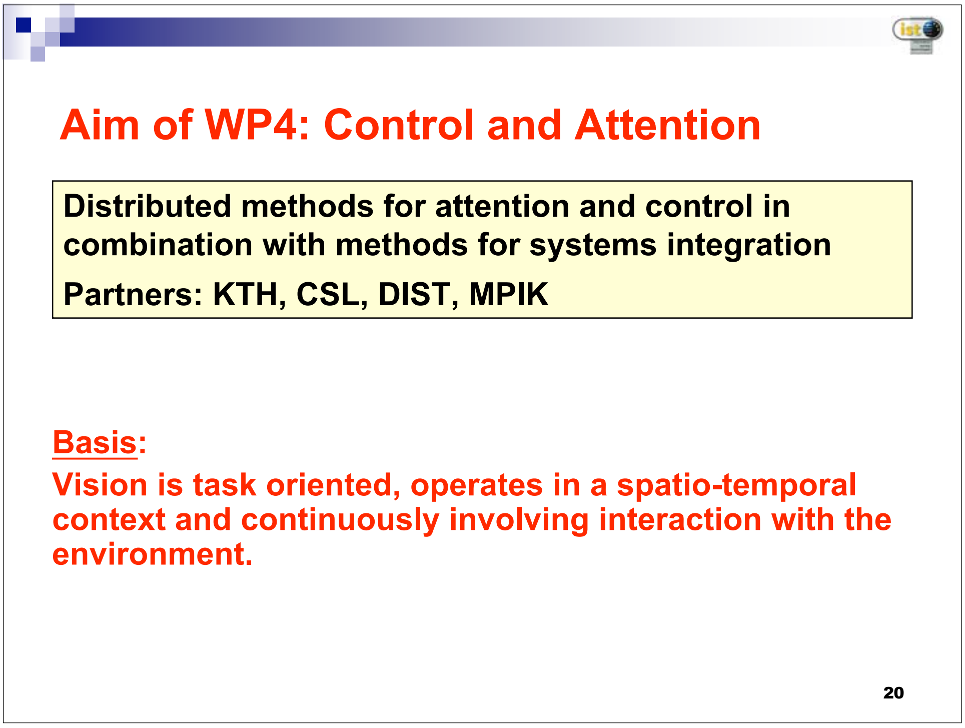
The algorithm is linear in the sequence length (exponential in number of objects).

Labelling accuracy is very significantly increased over frame-by-frame classification.



Conclusions

- Spatio-temporal continuity reasoning can significantly improve tracking reliability
- Probabilistic decisions for logically equivalent choices



Aim of WP4: Control and Attention

Distributed methods for attention and control in combination with methods for systems integration

Partners: KTH, CSL, DIST, MPIK

Basis:

Vision is task oriented, operates in a spatio-temporal context and continuously involving interaction with the environment.

Overview of WP4 Highlights

- **Combining proprioception and vision to learn about objects (DIST, MPIK)**
- **Coalitional game with salient peaks**

Learning Multimodal Object Models through Action

I. **Object Behavior Models**: demonstration of how the Behavior of an object when pushed/pulled/dropped can be used to characterize it

II. **Vision-Proprioception models:** a proposed way to combine proprioceptive & visual info, to build a novel VTMap representation - useful for action as well as speeding up recognition

III. **Tactile Models:** it has been shown how tactile information from grasping affordances can be used independently to distinguish objects.

Learning Object Models through Active Manipulation

Rao, Natale (DIST)

Wallraven (MPIK)

"How can Proprioception, Vision, & Active Control make object recognition more robust?"

Proprioceptive ViewTransition Map

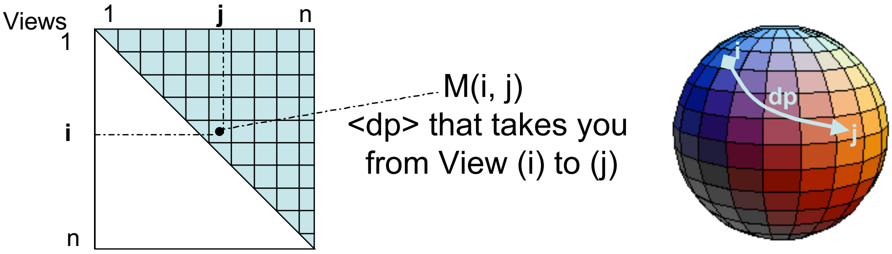
Self-terminating Exploration

Object Recognition

The Proprioceptive View-Transition Map: What is it?

An Object Representation that:

"Links Model Views in Proprioceptive Space"
(ako Proprioceptive viewing-sphere)



M(i, j)
<dp> that takes you from View (i) to (j)

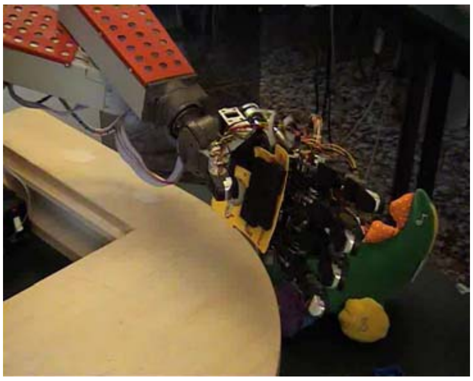
24

Learning Tactile Object Models through Grasping

(Lorenzo Natale, Ph.D. Thesis)

How can a robot grasp an unknown Object?

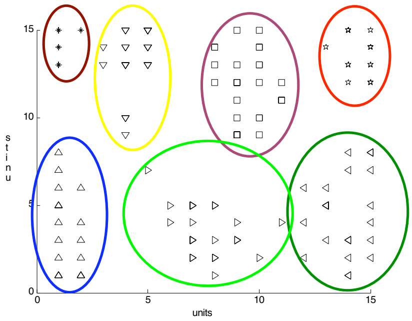
- **Use a simple motor synergy to flex the fingers and close the hand**
- **Exploit the intrinsic elasticity of the hand; fingers bend and adapt to shape of the object**



Result of Tactile Clustering

- **2D Self Organizing Map (100 neurons)**
- **input: proprioception (hand posture, touch sensors were not used)**

The SOM forms 7 classes (6 for the objects plus 1 for the no-object condition)



Conclusions

- **Visual object models can be learnt to include action knowledge:**
  - object behaviour
  - vision-proprioception maps
  - grasping affordances
- **Such models improve object recognition**

Coalitional game with salient peaks

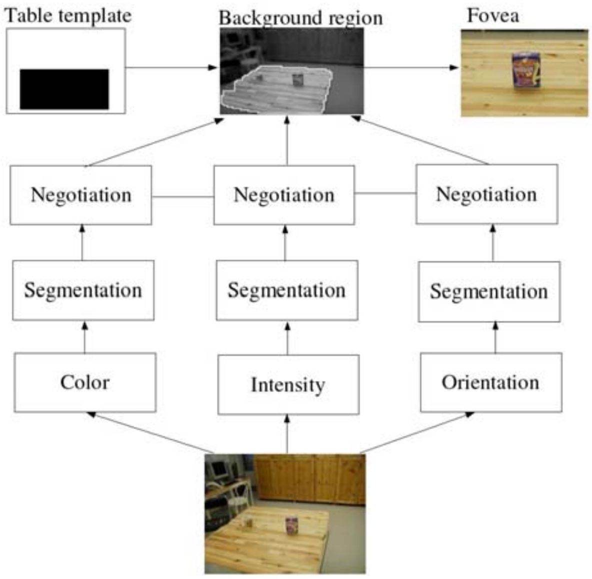
- **Salient peaks with respect to a set of different visual cues are computed.**
- **A set of selected peaks are represented by agents.**
- **Agents negotiate to form coalitions representing salient targets.**
- **Coalitions are formed when a stable set (von Neuman & Morgenstern 1944) can be achieved.**
- **Cue integration is tuned by a set of weights,**
  - **hence a task can be specified by tuning of weights.**

Example Scene

Coalitional game using redundant segmentation

- **An image is decomposed to a set of visual cues, which are processed at a distinct set of nodes.**
- **At each node a segmentation algorithm computes a set of segments at different scales.**
- **Segments negotiate to form coalitions.**
- **Configural and contextual information is used to guide an active foveated camera.**

Configural information

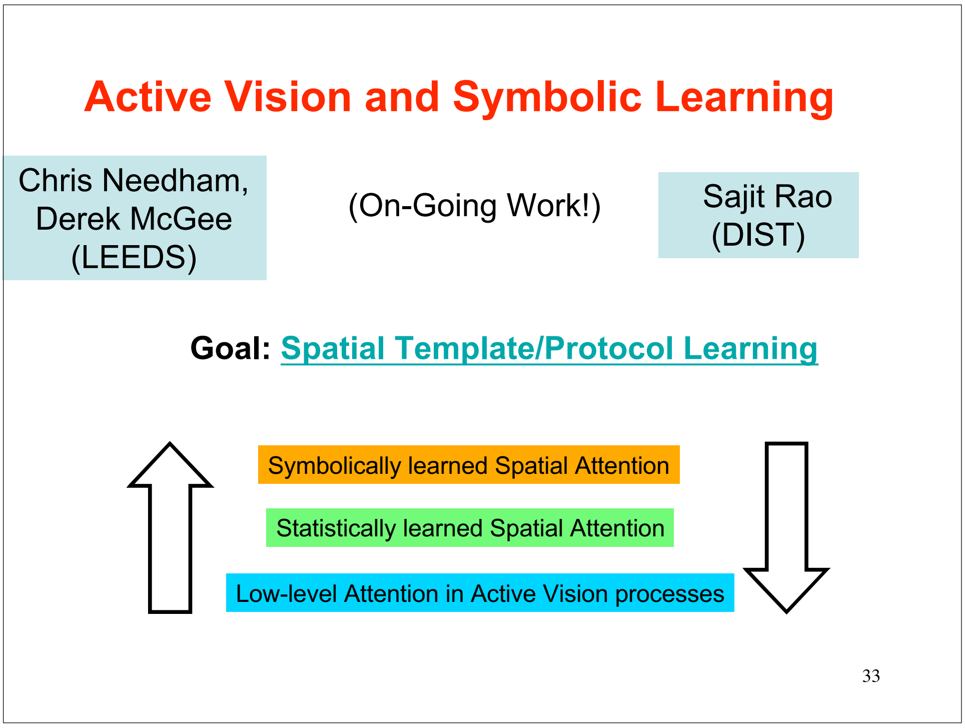


31

Conclusions

- **The attentional model enables *distributed processing* while avoiding bottlenecks such as data sharing and central control.**
- **The extraction of background regions enables *layout* understanding and *local context* extraction.**
  - Layout and local context are used to prune the number of irrelevant interest points and increase target saliency,
  - can potentially be used for other cognitive processes.

32



Active Vision and Symbolic Learning

Chris Needham,
Derek McGee
(LEEDS)

(On-Going Work!)

Sajit Rao
(DIST)

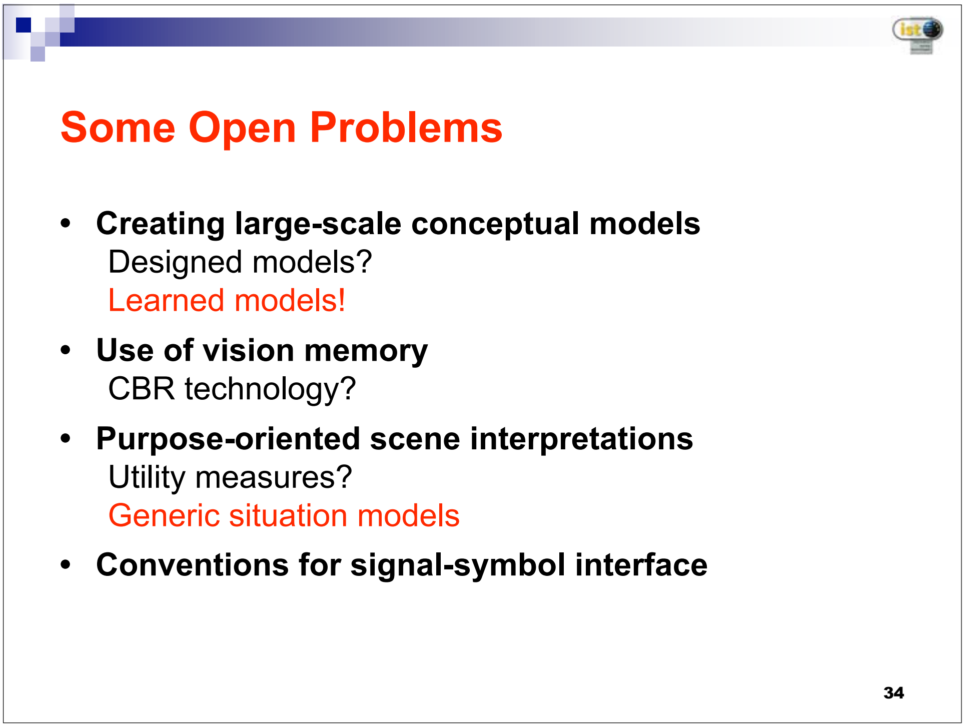
Goal: Spatial Template/Protocol Learning

Symbolically learned Spatial Attention

Statistically learned Spatial Attention

Low-level Attention in Active Vision processes

33



Some Open Problems

- **Creating large-scale conceptual models**
  Designed models?
  Learned models!
- **Use of vision memory**
  CBR technology?
- **Purpose-oriented scene interpretations**
  Utility measures?
  Generic situation models
- **Conventions for signal-symbol interface**

34